Linx
Revue des linguistes de l'université Paris X Nanterre

49 | 2003
L'actualité des notions d'interlangue et d'interaction exolingue

Appropriation du lexique lors d'un séjour linguistique : une étude de cas quantitative et qualitative

Maarit Mutta

Édition électronique
URL : http://journals.openedition.org/linx/555
DOI : 10.4000/linx.555
ISSN : 2118-9692

Éditeur
Presses universitaires de Paris Nanterre

Édition imprimée
Date de publication : 1 décembre 2003
Pagination : 109-123
ISSN : 0246-8743

Référence électronique
Maarit Mutta, « Appropriation du lexique lors d'un séjour linguistique : une étude de cas quantitative et qualitative », *Linx* [En ligne], 49 | 2003, mis en ligne le 17 mars 2011, consulté le 21 avril 2019. URL : http://journals.openedition.org/linx/555 ; DOI : 10.4000/linx.555

Département de Sciences du langage, Université Paris Ouest

Appropriation du lexique lors d'un séjour linguistique : une étude de cas quantitative et qualitative

Maarit MUTTA
Université de Turku

Toute recherche linguistique requiert une réflexion approfondie quant aux choix méthodologiques concernant un contexte spécifique. Dans le cadre de l'appropriation des langues (apprentissage et acquisition), le chercheur se demande, entre autres, quelle compétence langagière étudier, quel type d'analyse effectuer, quel corpus choisir, quelles méthodes adopter pour y arriver et comment exposer les résultats. Nous aborderons cette problématique du point de vue du choix de la méthode d'analyse, à savoir qualitative ou quantitative, sur un corpus spécifique, en examinant l'impact d'un séjour linguistique dans une université française (à l'Université de Provence - Aix-Marseille I) sur la compétence lexicale d'étudiants universitaires finlandais (de l'Université de Turku).

Pourquoi avons-nous eu recours à une analyse quantitative *et* qualitative ? La recherche empirique sur l'appropriation des langues utilise le plus souvent des analyses qualitatives. L'utilisation d'une analyse quantitative n'est pas exclue, mais fait souvent l'objet de controverses. L'opposition à l'approche quantitative concerne tout le domaine des sciences humaines, dans la mesure où celles-ci ne sont pas considérées comme des sciences exactes (Suomela-Salmi 1997 : 16-17). Cependant, des recherches antérieures montrent l'utilité des mesures d'ordre objectif, c'est-à-dire quantitatif, dans diverses études de linguistique appliquée. A. Cossette (1994 : 1) affirme que la richesse lexicale, qui est normalement calculée d'après des formules statistiques, est une notion importante en didactique. Preston et Bailey (1996 : xiv), pour leur part, soulignent les avantages de l'analyse quantitative lorsqu'il s'agit d'étudier la variation interlinguale, en relation avec le point de vue sociolinguistique. Tarone (1988 : 131) fait remarquer que les résultats d'une étude qualitative sur l'interlangue devraient être étayés par des études quantitatives afin de pouvoir déterminer la généralité des résultats obtenus auprès d'apprenants d'une langue seconde. Il est à noter que les études quantitatives requièrent, en général, des échantillons suffisamment larges pour qu'on puisse tirer des conclusions définitives sur les phénomènes à étudier, c'est-à-dire pour déterminer

s'ils sont statistiquement significatifs. Pour Uusipaikka, statisticien[1], ce postulat de généralité s'avère juste, mais il ajoute qu'avant de recueillir un corpus très vaste, il faut en définir en détail les prémisses, c'est-à-dire la justification des méthodes et des procédures. Certes, la systématicité du recueil alliée à la taille du corpus augmente la fiabilité des résultats, mais le gain en fiabilité ne s'élève pas à la puissance linéaire (cf. le seuil de saturation).

Dans ce qui suit, nous traiterons de certaines possibilités et de certaines limites des analyses quantitative et qualitative à partir de l'expérience de notre recherche, dont le but était de donner une image plus précise et plus complète de la richesse lexicale des étudiants et de son développement après un séjour linguistique à l'étranger.

1. L'interlangue dans une perspective longitudinale

Nous n'entrerons pas ici dans une discussion détaillée sur les origines et les caractéristiques du phénomène nommé interlangue (IL). Pourtant, il vaut la peine de rappeler que ce concept, issu des recherches de Selinker (1972/1975), n'est pas exempt de critiques. Selon Porquier, la notion a rapidement évolué, ayant été souvent associée initialement à l'apprentissage guidé et élargissant ensuite son champ d'investigation à l'acquisition naturelle des langues, ce qui a fait apparaître de nouvelles notions décrivant mieux le phénomène (1986 : 101-107). Giacobbe signale à son tour « le manque d'homogénéité des différentes définitions du terme 'interlangue' » (1992 : 24). Nous avons cependant adopté le terme d'interlangue, devenu le plus fréquemment utilisé dans la littérature spécialisée, indépendamment de la terminologie utilisée par d'autres auteurs[2]. Les IL sont caractérisées par leur variabilité, leur systématicité et leur dynamique évolutive. Il s'agit de systèmes structurés, mais en même temps en évolution constante ; en d'autres termes, elles sont instables, se transformant au fur et à mesure qu'elles se rapprochent tendanciellement de la langue-cible, donc d'une compétence native (cf. Faerch *et al.* 1984).

Nos deux corpus (le corpus primaire et celui du groupe de contrôle) ayant été recueillis à deux moments différents (avant et après un séjour linguistique, cf. *infra*), nous aurions pu les examiner d'un simple point de vue synchronique, mais notre objet de recherche aurait alors plutôt été la performance (cf. Elo 1993). Or, pour pouvoir examiner la compétence des étudiants, il faut recourir à une étude développementale. C'est pourquoi nous avons cherché à constater des progrès dans le temps et éventuellement à détecter des formes stabilisées des IL[3].

---

[1] Cours intitulé « Langue, homme et technologie » (*Kieli, ihminen ja teknologia*) portant sur le rôle des statistiques dans la recherche linguistique, lors du séminaire du 10 mai 2002 à l'université de Turku.

[2] Parmi d'autres, K. Vogel (1995) présente le phénomène de l'interlangue d'une manière quasi exhaustive ; entre autres, il le restitue parmi les divers courants et les modèles relatifs à l'acquisition des langues secondes. Vogel considère donc la recherche sur l'apprentissage des langues comme une recherche interdisciplinaire : elle se doit d'intégrer les approches d'autres disciplines scientifiques, par ex. la psychologie, la linguistique et les sciences de l'éducation.

[3] Gaonac'h considère qu'une analyse synchronique pure revient en pareil cas à appauvrir la notion d'interlangue, car elle sort du cadre évolutif de l'acquisition des systèmes successifs orientés vers la compétence en L2. Du point de vue pédagogique, il propose de plus une hypothèse sous-jacente à

Le corpus d'ensemble de notre recherche (Mutta 1999) est présenté dans le Tableau 1 :

Examen des connaissances langagières Dissertations (environs 150-200 mots)				
Corpus	niveau I	séjour linguistique	niveau II	Total
groupe primaire/ d'Aix	53	1 ou 2 semestres à l'Université de Provence - Aix-Marseille I	53	106
groupe de contrôle/ témoin	25	entre 4 semaines et 18 mois dans un pays francophone ; pas d'études universitaires	25	50

Tableau 1. Corpus

Le corpus est composé de 106 (2x53) dissertations écrites recueillies en deux phases différentes lors d'une épreuve de connaissances langagières[4], et de 50 (2x25) dissertations formant le groupe de contrôle. Les étudiants du premier groupe ont rédigé leur première dissertation avant d'effectuer un séjour linguistique d'un ou deux semestres en France dans le cadre d'un échange ERASMUS/SOCRATES. Les étudiants du second groupe (groupe de contrôle), quant à eux, n'ont pas participé à ce genre d'échange, mais ont néanmoins séjourné dans un pays francophone. Les étudiants des deux groupes ont rédigé la seconde dissertation après leurs séjours respectifs à l'étranger.

Quant à l'environnement linguistique, notre corpus peut être considéré comme étant situé entre les deux extrêmes ; nous pouvons parler de milieu mixte, c'est-à-dire d'appropriation dans un contexte semi-institutionnel. Même si les étudiants du corpus primaire suivaient un enseignement universitaire, il ne s'agissait cependant pas d'un cours spécial de langue, mais de cours destinés aux étudiants (majoritairement francophones) autochtones. De ce fait, les étudiants du groupe primaire ont eu accès au moins à trois types d'input linguistique en français : a) l'enseignement du français à l'Université de Turku, b) l'appropriation naturelle (par exemple. par des contacts personnels) en France et c) l'enseignement semi-guidé à l'université. De leur côté, les étudiants du groupe de contrôle ont eu accès à deux ou trois types d'input : a) l'instruction en Finlande, b) des contacts personnels et/ou c) quelques enseignements (par ex. l'assistance à des cours de langue type « Alliance française ») mais d'une durée brève, lors du séjour en France. Une de nos hypothèses était que le séjour dans un cadre universitaire améliorerait les compétences écrites et surtout les

ces développements : « l'élève a une certaine connaissance […] de 'son' interlangue à chaque moment de son apprentissage, et il faut lui permettre d'en tirer le meilleur parti » (1984 : 66, 77).

[4] Les étudiants passent l'examen deux fois durant leurs études ; d'abord au niveau des études de base (à la fin de leur première année d'études ou pendant le premier semestre de leur seconde année), et ensuite à la fin de leurs études de licence (Mutta 1999 : 5).

connaissances lexicales plus efficacement — c'est-à-dire quantitativement et qualitativement — qu'un séjour linguistique dans un milieu non universitaire. Deux autres hypothèses ont été également faites sur la base d'une étude préalable : d'abord, que la production deviendrait plus naturelle après le séjour linguistique (par exemple quant à l'utilisation des organisateurs pragmatiques ou textuels cf. *infra*) et ensuite, que des traits du langage parlé pourraient apparaître dans la production écrite. Nous ne traiterons pas de ce dernier point dans cet article (cf. Mutta à paraître).

2. L'analyse quantitative

Si notre recherche inclut une analyse quantitative et qualitative, la dimension quantitative s'y trouve néanmoins légèrement plus développée. L'analyse quantitative s'est faite en deux volets : d'abord l'étude des facteurs linguistiques quantitatifs, et ensuite celle des corrélations entre différents facteurs et différents groupes. Le Tableau 2 présente cette procédure :

Méthode de M. Linnarud : *Lexis in composition* (1986)	
• la variation lexicale	= le pourcentage des vocables lexicaux / toutes les occurrences lexicales
• la densité lexicale	= le pourcentage des mots lexicaux et grammaticaux/ nombre total de mots
• l'individualité lexicale	= les mots à occurrence unique chez un étudiant dans tout le corpus
• la sophistication lexicale	= le degré de difficulté

Les corrélations entre les facteurs linguistiques : (LOGICIEL (SPSS))
• la corrélation des valeurs, ou coefficient de Bravais-Pearson (r)
• la corrélation des rangs, ou coefficient de Spearman (ρ) ou coefficient de Kendall
• un t-test et un test *ANOVA*

Tableau 2. Méthode quantitative

La méthode quantitative est essentiellement basée, sous une forme adaptée, sur la méthode utilisée par Linnarud dans son ouvrage ***Lexis in composition*** (1986). Afin d'effectuer nos calculs, nous avons d'abord eu recours au logiciel ***WordCruncher***[5], qui calcule chaque élément entre deux blancs comme mot isolé ; il s'ensuit que le concept 'mot' est défini d'une façon orthographique, sauf quelques unités étroitement liées l'une à l'autre que nous avons estimées difficiles à séparer (par ex. ***n'est-ce pas, parce que, d'après***). En outre, les calculs ont été faits à partir des formes non lemmatisées (voir entre autres Gustafsson 1990 : 48-49). Après avoir défini le mot, nous avons

[5] Logiciel développé à l'université de Brigham Young, aux Etats-Unis, pour la recherche lexicale. Une fois que le texte à examiner est introduit dans le programme, le mot peut être examiné dans son contexte linguistique, et le logiciel permet d'établir des listes de concordance.

poursuivi notre analyse concernant les facteurs linguistiques quantitatifs, à savoir la variation lexicale, la densité lexicale, l'individualité lexicale et la sophistication lexicale. On suppose que l'ensemble de ces phénomènes donne une image plus complète de la richesse lexicale de chaque étudiant que l'analyse d'un phénomène isolé.

Dans cette recherche, la variation lexicale (VL) comprend le pourcentage des vocables lexicaux de toutes les occurrences lexicales[6] (en anglais ***type/token ratio***), tandis que la densité lexicale (DL) indique le pourcentage des mots lexicaux par rapport au nombre total des mots dans chaque dissertation. Les valeurs de l'individualité (ILe) et de la sophistication lexicales se mêlent et s'influencent l'une l'autre. De son côté, l'individualité correspond à l'originalité lexicale calculée d'après les mots à occurrence unique (***hapax legomena***)[7] chez un étudiant par comparaison avec le nombre de mots lexicaux dans la dissertation en question, tandis que la sophistication comprend le degré de difficulté[8] ; selon des calculs, plus le texte est original, plus le niveau de sa sophistication est élevée (Linnarud 1986 ; Mutta 1999). Cette dernière valeur recoupe d'une certaine manière l'analyse qualitative.

Deuxièmement, afin de voir si les facteurs linguistiques divers des deux corpus et leurs changements éventuels d'un niveau à l'autre sont statistiquement significatifs, nous avons eu également recours au logiciel ***SPSS*** (***8.0 for Windows***) pour calculer des corrélations entre différents facteurs, niveaux et corpus. Les corrélations calculées à l'aide du logiciel sont les suivantes : corrélation des valeurs (coefficient de Bravais-Pearson, ***r***), corrélation des rangs (coefficient de Spearman, ρ et coefficient de Kendall, ***Kendall's tau-b***), t-test et test ANOVA[9].

Avant d'entamer la discussion sur les possibilités et les limites d'une analyse quantitative dans ce genre de recherche, nous expliciterons les modalités de la deuxième partie de notre analyse.

---

[6] Normalement (par exemple dans les études de cas), le pourcentage des vocables et des occurrences concerne tous les mots, soit mots pleins soit mots outils, mais dans notre recherche, à l'instar de Linnarud, nous nous sommes intéressée seulement à la variation des mots lexicaux ou pleins.

[7] Du fait du nombre de formes erronées, nous avons rectifié les occurrences avant certains calculs (*cf.* aussi Mutta, à paraître).

[8] Cette valeur a été calculée manuellement en comparant le vocabulaire des apprenants, notamment les mots à occurrence unique, avec les fréquences indiquées dans le *Dictionnaire des fréquences* (1971) et le *Français Fondamental* (1972) ; ce dernier était à l'époque l'outil de base pour la sélection du vocabulaire français à apprendre dans les lycées finlandais (Mutta 1999 : 48). De plus, nous avons comparé les 20 mots les plus fréquents utilisés dans le corpus avec les listes du *Dictionnaire des fréquences*.

[9] Nous sommes reconnaissante à K. Lertola, statisticien, qui a effectué les calculs du t-test et du test ANOVA ; il a également calculé une partie des corrélations, et nous-même le reste.

3. L'analyse qualitative

L'analyse quantitative seule aurait donné des résultats montrant des tendances générales du corpus. Mais ces résultats, notamment les corrélations, n'auraient révélé que le rapport entre divers facteurs sans rien dire de la nature des changements. Pour expliquer ces derniers, une approche qualitative y a donc été associée.

Dans cette partie de l'analyse, nous avons traité des mots du vocabulaire des apprenants dans leur contexte linguistique et non comme items séparés. L'analyse était bipartite : d'abord, nous avons isolé des erreurs ou formes déviantes d'une façon assez générale sans en faire d'étude minutieuse au sens de Corder (*cf.* Corder 1980) ; la reconnaissance, la description et l'explication des formes idiosyncrasiques étaient vues selon la perspective de la parole en jugeant leur impact sur la lisibilité (*cf.* Péry-Woodley 1993). Ensuite, nous avons examiné les dissertations du point de vue de la cohérence textuelle, en tant que contribuant à la compréhension du texte écrit, plus précisément à la clarté et à la lisibilité du texte. En conséquence, les organisateurs pragmatiques sont définis, premièrement, comme des moyens de ponctuer le texte ou de le découper en paragraphes, et deuxièmement, comme des marqueurs textuels explicites jouant un rôle de 'passeurs' d'une phrase à l'autre.

A première vue, la ponctuation semble un phénomène peu pertinent quant à la cohérence textuelle. Cependant Fayol (1989 : 37), à maintes reprises, a pu montrer que la ponctuation représente une sorte de modèle mental sous-jacent du texte que le sujet écrivant émet en organisant le texte. Bessonnat (1988 : 94) va dans le même sens en discutant du découpage d'un texte en paragraphes, ce qui constitue selon lui un outil métatextuel. Il estime que le découpage en paragraphes a un « triple rôle » : un rôle de facilitation de la lecture, un rôle de programmation hiérarchique du texte et un rôle dialogique (*op. cit.* : 85-87). Les autres organisateurs pragmatiques correspondent dans cette étude aux connecteurs et aux organisateurs textuels classifiés principalement par Adam (1990), par Elo (1993), par Lundquist (1980, 1983), et par Riegel *et al.* (1997). Dans le Tableau 3 ci-dessous sont présentés de manière regroupée les termes utilisés. L'analyse des organisateurs est liée à l'hypothèse que la production des étudiants deviendrait plus naturelle après le séjour linguistique à l'étranger.

ORGANISATEUR		EXEMPLES	CATÉGORISATION ORIGINALE
Additif		et, encore, de plus	Lundquist
Enumératif		d'abord – ensuite – enfin	Lundquist
Transitif		d'ailleurs, d'autre part	Lundquist
Explicatif/ reformulatif		car, à savoir, en d'autres termes	Adam Lundquist
Illustratif		par exemple, notamment	Lundquist
Comparatif		plutôt	Lundquist
Adversatif		or, mais, en revanche	Lundquist
Concessif		néanmoins	Lundquist
Causatif/ consécutif/ conclusif		donc, ainsi, en effet	Lundquist
Résumatif/ reformulatif		bref, en somme	Adam Lundquist
Présentatif		c'est, il y a	Lundquist
Organisateur temporel		puis, en même temps	Adam Lundquist Riegel *et al.*
Organisateur spatial		ici, là	Adam, Riegel *et al.*
Connecteur argumentatif		mais, même si	Adam Riegel *et al.*
Marqueur de clôture/ d'ouverture		voilà	Adam
Marqueur métatextuel	Modalisateur argumentatif	évidemment, justement, même	Riegel *et al.*
	Commentaire métatextuel	voir p., cf.	Lundquist
	Introducteur	je crois/ pense/ trouve que	Elo

Tableau 3. Catégories des organisateurs pragmatiques

4. Bilan des analyses

Les résultats statistiques montrent, dans l'ensemble du corpus, c'est-à-dire dans le corpus primaire et le corpus témoin, une différence statistiquement significative dans l'évolution du niveau I (avant le séjour linguistique) au niveau II (après le séjour linguistique), quant à certains facteurs quantitatifs, notamment la valeur du nombre de mots et des "mots uniques" (*cf.* la sophistication lexicale). Les tableaux 4 et 5 présentent ces corrélations ; outre les facteurs linguistiques quantitatifs, le nombre de mots a été calculé. Dans ces tableaux, ILe (a) correspond à tous les mots à occurrence unique, tandis que ILe (b) correspond aux mots rectifiés à occurrence unique (*cf.* Mutta 1999 et Mutta à paraître).

	mots I/II	VL I/II	DL I/II	ILe a I/II	ILe b I/II
mots I/II					
VL I/II	/-.322			.563/.528	.511/.559
DL I/II					
ILe a I/II					.876/.837
ILe b I/II					

Tableau 4. Coefficients significatifs de Bravais-Pearson aux deux niveaux

	mots I/II	VL I/II	DL I/II	ILe a I/II	ILe b I/II
mots I/II					
VL I/II	/-.182			.405/.336	.349/.380
DL I/II					
ILe a I/II					.695/.584
ILe b I/II	.160/				

Tableau 5. Coefficients significatifs de Kendall aux deux niveaux

Par ailleurs, le test ANOVA[10], qui a permis d'effectuer les calculs comparatifs entre les deux corpus, montre qu'il n'y a pas de différence significative entre ceux-ci quant aux changements de ces facteurs linguistiques quantitatifs. Nous en avons conclu que les changements entre les deux corpus étaient presque identiques. Les corrélations calculées entre les différents facteurs quantitatifs, séparément aux deux niveaux, indiquent néanmoins que la variation lexicale semble bien corréler les autres facteurs. Nous avons conclu de ces corrélations que les valeurs de la variation lexicale et des "mots uniques" étaient des facteurs quantitatifs centraux dans les calculs, et que la quantité de "mots uniques" était davantage liée à la valeur VL qu'à la valeur du nombre de mots. Autrement dit, plus la variation lexicale augmente, plus il y a de mots uniques, et par conséquent le degré de difficulté augmente, ce qui, de son côté, est un indice de la croissance de la richesse lexicale. Ces données statistiques infirment en partie notre première hypothèse, qui postulait que le séjour linguistique dans un cadre universitaire améliorerait les connaissances lexicales plus efficacement qu'un séjour non universitaire.

L'analyse de la sophistication lexicale, qui recoupe en partie l'analyse qualitative, s'est avérée intéressante : nous avons comparé les 20 mots les plus fréquents dans les deux corpus avec le Dictionnaire des fréquences (1971)[11]. Il en ressort qu'après le séjour linguistique (niveau II), les onze premiers mots sont les mêmes dans les deux corpus et dans le Dictionnaire des fréquences, mais que leur rang n'est pas tout à fait le même, sauf pour le premier (la préposition *de*). Les mots et leur fréquence dans les deux corpus sont présentés dans le Tableau 6 à la page suivante.

Ainsi, les deux corpus respectent assez bien les tendances générales concernant l'utilisation des mots fréquents, en particulier le corpus primaire. Nous avons donc pu

[10] Pour les calculs exacts, voir Mutta (à paraître).

[11] Nous avons utilisé des formes lemmatisées afin de comparer les fréquences. Dans la lemmatisation, nous avons adopté principalement la même pratique que dans le *Dictionnaire des fréquences* : entre autres, tous les verbes sont lemmatisés sous la forme de l'infinitif ; en revanche, par exemple, les articles ne sont pas lemmatisés.

conclure que plus il y avait de mots dans le corpus, mieux les mots étaient distribués de façon comparable à la tendance générale des textes français. Autrement dit, les dissertations des étudiants se rapprochent globalement des textes de scripteurs natifs quant à l'utilisation des mots les plus fréquents, qui sont, comme nous pouvons le voir dans le Tableau 6, des mots grammaticaux[12]. Le rang élevé du verbe *avoir* chez les apprenants est probablement lié à sa fonction d'auxiliaire ou à sa disponibilité dans leur vocabulaire. Nous ne commentons pas ici plus en détail les autres mots et rangs. De ce point de vue, les textes sont devenus plus naturels, surtout dans le corpus primaire, comme nous le supposions dans la deuxième hypothèse.

corpus d'Aix mot / fréquence			dictionnaire des fréquences	corpus témoin mot / fréquence		
niveau I	niveau II	I&II		niveau I	niveau II	I&II
être 485	de 570	de 1028	1. de (de + d')	de 246	de 267	de 513
de 458	être 418	être 903	2. la	être 199	être 256	être 455
avoir 328	la 355	la 648	3. être	avoir 169	avoir 178	avoir 347
les 297	avoir 309	avoir 637	4. et	les 156	les 158	les 314
la 293	le 266	les 506	5. que (que + qu')	la 143	et 143	la 279
et 222	et 226	le 448	6. le	que 116	la 136	et 254
que 212	que 218	et 448	7. à	et 111	que 133	que 249
à 1 95	à 214	que 430	8. l'	l' 94	le 125	le 211
l' 182	les 209	à 409	9. avoir	à 87	l' 108	l' 202
le 182	il 188	il 356	10. les	le 86	il 93	à 173
on 170	l' 172	l' 354	11. il	il 78	à 86	il 171
il 168	des 162	des 327	12. ne	des 76	des 84	des 160
des 165	en 147	en 306	13. je	pas 73	on 83	pas 155
en 159	on 133	on 303	14. un	on 72	pas 82	on 155
pas 155	pas 130	pas 285	15. se	je 66	je 78	je 144
un 128	un 121	un 249	16. des	qui 52	un 76	en 144
mais 107	une 116	qui 217	17. en	un 52	en 71	un 128
qui 106	qui 111	une 208	18. qui	plus 50	plus 63	plus 113
je 99	dans 109	je 207	19. une	n' 48	mais 62	mais 105
ne 99	je 108	dans 198	20. dans	ne 47	j' 58	qui 102
						j' 102
						ne 102
			23. pas			
			28. plus			
			30. on			
			34. mais			
total 11.069	11.332	22.401		total 5.368	5.988	11.356

Tableau 6. Mots les plus fréquents

[12] Cela ressemble à un phénomène de la théorie du chaos appliquée à l'acquisition des langues secondes, entre autres par D. Larsen-Freeman (*Chaos/Complexity Theory*). A l'instar de Schroeder (1995), Larsen-Freeman explicite la loi de Zipf connectant le rang et la fréquence des mots : « In other words, if a word occupies a particular word frequency rank in a given language, then it is likely to reflect that same frequency *in any given text* of that language. Zipf's law is apparently not only applicable to a language in general, but also to specific writers. » (1997 : 150). Les déviances dans le rang peuvent être liées à l'utilisation erronée de différentes formes.

La partie qualitative de l'analyse concernant les formes erronées montre l'impossibilité de mettre en évidence de grandes différences entre les deux corpus quant aux erreurs lexicales dans la production. Nous avons supposé qu'il s'agissait de variabilités individuelles. Cependant, nous avons pu distinguer un développement général du niveau I au niveau II, la correction lexicale devenant meilleure, malgré des fautes d'orthographe (surtout sur les accents).

En ce qui concerne l'utilisation des organisateurs pragmatiques, il ressort de l'analyse qualitative que les résultats du corpus primaire sont meilleurs que ceux du groupe de contrôle ; les étudiants utilisent par exemple une variété plus importante d'organisateurs pragmatiques, soit comme des moyens de ponctuer le texte ou de le découper en paragraphes, soit comme des marqueurs textuels explicites (par exemple. des connecteurs adversatifs *mais*, *pourtant*, ou conclusifs *donc*), mais avec beaucoup de variabilités. Les résultats du corpus témoin semblent donc aller dans le même sens que ceux du corpus primaire — sauf que la variété des organisateurs semble moindre – infirmant ainsi en partie notre hypothèse sur les différences entre les corpus. De toute façon, nous avons pu conclure que les capacités métatextuelles de tous les apprenants s'amélioraient comme prévu pendant le séjour linguistique à l'étranger (*cf.* néanmoins la discussion dans Mutta à paraître). L'hypothèse selon laquelle la production devient plus naturelle, entre autres, grâce à l'augmentation des organisateurs pragmatiques se trouve ainsi vérifiée.

Quant au développement de la cohérence textuelle, nous avons discerné dans l'ensemble une cohérence plus élevée au niveau II qu'au niveau I. Nous l'avons constaté clairement dans les dissertations des étudiants de matière secondaire. Au niveau I (avant le séjour), les apprenants utilisaient beaucoup de phrases isolées ; ou alors les moyens linguistiques utilisés pour contribuer à la cohérence du texte n'étaient présents qu'à l'intérieur des paragraphes qui eux, restaient isolés. Au niveau II (après le séjour), au contraire, l'enchaînement était plus cohérent non plus seulement au niveau intra- et interphrastique ; mais également entre les paragraphes. En voici un exemple (fidèlement retranscrit ici) :

> Ex 1. niveau I :
> Si on compare le nombre d'étrangers en Finlande avec ce des autres pays nordiques par. ex la Suède et le Danemark, on peut bien voir que le nombre est petit. A mon avis, il n'y en a pas trop. Si la Finlande ne rejoindra pas la C.E.E., notre pays sera trop extérieur d'Europe. Les étrangers ne me font pas peur même s'il y en avait beaucoup. Les étrangers peuvent nous apprendre de ce qu'ils savent bien, des difficultés sont été crées pour les resoudre. Je pense que les réfugiés ont seulement deux possibilités : souvent c'est d'être tortu ou même mourir en son propre pays et la deuxième possibilité c'est d'immigrer dans un pays étrange. Je ne connais pas bien le nombre exact de réfugiés en Finlande mais il y en a pas trop. C'est souvent les préjugés des gens qui constituent des obstacles à l'intégration des étrangers. Je ne connais pas personnellement des étrangers en Finlande mais au contraire j'ai des amis en étranger.

> Ex. 2. niveau II : Après avoir passé quelques mois à Aix-en-Provence, j'ai commencé à réfléchir si je veux habiter en Finlande pour l'éternité. En Finlande je suis attaché à mes amis. Mon loisir et mes études me plaisent et la variation des saisons. Mais en ce qui concerne la vie culturelle, il y a quelque chose qui me manque. A Aix, j'aimais bien m'asseoir dans les cafés avec mes amis. C'était magnifique de passer le temps en regardant les passants, en regardant la vie animé et en parlant avec des chalereux méridionaux. C'est triste qu'en Finlande il nous manque de bons cafés et que dans les plusieurs cafés on doit se servir soi-même ! En plus, on ne peut pas voir des films français ici autant qu'en France. Cela me manque aussi. Je ne sais pas, peut-être que je me trouverai à l'étranger après les études !

5. Conclusion en forme de perspectives

Les tendances générales observées dans l'ensemble du corpus témoignent, entre autres, de l'amélioration de la richesse lexicale et de l'augmentation de la cohérence textuelle après le séjour linguistique en pays francophone. Cela est une preuve, au moins partielle, de l'avantage de ce mode d'appropriation dans un milieu naturel, dû à l'accès à un input plus varié que celui offert dans le milieu guidé en Finlande. Intuitivement, d'après la première lecture du matériel, nous étions d'avis qu'il y aurait une différence entre les deux corpus, que nous avons essayé de révéler au moyen de notre méthode d'analyse.

Les problèmes méthodologiques principaux rencontrés pendant la recherche étaient les suivants :

– les critères pour recueillir les deux corpus
– le vocabulaire en tant que centre d'intérêt
– la définition des différentes catégories à examiner.

Si l'on examine le premier point, nous pouvons constater qu'il nous manquait une sorte d'homogénéité dans les critères de collecte. Cela était dû aux moments de collecte qui, de leur côté, étaient déterminés par la taille du corpus. Afin d'avoir un corpus suffisamment étendu pour pouvoir tirer quelques conclusions sur les tendances générales (une analyse quantitative, donc), nous avions décidé de recueillir environ 50 dissertations d'étudiants pour le corpus primaire. L'étude longitudinale est primordiale pour pouvoir examiner le progrès dans les compétences des apprenants, mais en tant que moyen de recherche, elle est assez lourde à réaliser d'une manière idéale, au moins si l'objet de recherche est soumis à une analyse quantitative, c'est-à-dire sur un certain nombre de cas. En ce qui concerne le choix du corpus témoin, les critères étaient les mêmes, mais selon un choix aléatoire. Pour améliorer une telle approche, il conviendrait premièrement de faire passer le même test à tous les sujets, et ensuite d'y ajouter un questionnaire ou un entretien concernant, entre autres, les connaissances linguistiques des apprenants. Cette procédure faciliterait l'interprétation des analyses.

Il va de soi que le lexique, en tant qu'objet d'une étude qualitative, semble plus difficile à maîtriser que par exemple les structures grammaticales car c'est une catégorie « ouverte ». Nos problèmes étaient liés à trop de variables à examiner et à

trop de variabilités personnelles dans les résultats, surtout dans le corpus témoin. Pour résoudre ou réduire ce type de difficultés, on pourrait diminuer le nombre de catégories à examiner et regrouper les cas individuels d'après les caractéristiques les plus saillantes, de façon à constituer des groupes représentatifs qui, à titre d'exemple, dévoileraient quelques tendances éventuelles (*cf.* la méthodologie d'Arnaud 1984, 1993).

Quant au troisième problème évoqué, concernant la distinction entre mots pleins et mots outils dans le calcul des facteurs linguistiques quantitatifs, il conviendrait d'ajuster à nouveau les méthodes aux objectifs. Nous avions pour but de montrer que — et en quoi — les connaissances lexicales du corpus primaire surpasseraient en quantité et en qualité celles du groupe de contrôle après le séjour dans une université française, par comparaison avec un séjour non universitaire, selon la présomption d'un input linguistique plus grand et plus varié dans un tel milieu. Or, les calculs statistiques n'ont pas révélé là de différence significative entre les deux corpus.

Nous aurions pu définir les mots différemment, selon la définition orthographique certes, mais en regroupant les organisateurs pragmatiques comme une catégorie à part, pour avoir ainsi trois catégories différentes, ce qui aurait pu permettre de mieux expliciter les différences entre les deux niveaux et entre les deux corpus. Nous aurions également pu calculer les facteurs linguistiques quantitatifs à partir de tous les mots et pas seulement, à l'instar de Linnarud, à partir des mots lexicaux. Pourquoi n'avons-nous pas effectué une deuxième analyse quantitative avant de passer à l'analyse qualitative ? A titre d'information, le corpus principal consistait en 22.401 mots (ou occurrences), tandis que le groupe de contrôle consistait en 11.356 mots ; ainsi, la totalité des mots de nos deux corpus était de 33.757, répartis manuellement en mots pleins et mots outils avant les calculs quantitatifs. La grande taille du corpus explique pourquoi nous n'avons pas entrepris une nouvelle analyse à ce point de la recherche. En outre, nous avions effectué un test pilote en ce qui concerne la définition du concept 'mot' ; en résumé, nous avons pu constater d'après le test pilote que les résultats montrent les mêmes tendances en dépit de définitions différentes. Par conséquent, le plus pertinent était d'utiliser les mêmes critères d'un bout à l'autre du corpus.

En fin de compte, une analyse quantitative permet de traiter un grand nombre de mots, souvent à l'aide d'un logiciel qui, de son côté, permet de montrer les tendances générales dans un corpus, et de plus si celles-ci sont statistiquement significatives. Ce que l'analyse quantitative n'indique pas, c'est le rapport de cause à effet ; entre autres, les corrélations constatées ne fournissent pas d'explication sans interprétation. D'où l'utilité du recours à une analyse qualitative en complément d'une analyse quantitative. En général, ce genre d'analyse se fait sur un corpus de moindre taille, ce qui permet d'expliciter les phénomènes plus en détail ; pour un corpus plus vaste, un ajustement nécessaire sera requis.

Ce que nous venons de constater rejoint l'idée de Mäkelä (1990 : 59) : « Malgré des similarités importantes [entre les deux démarches], l'analyse qualitative est presque forcément plus individuelle et moins standardisée que le traitement du matériel

quantitatif »[13]. Il ajoute que pour tenir compte du critère de fiabilité (op. cit. : 47) [l'un des critères docimologiques de l'évaluation], il faut rendre les opérations techniques et celles de la pensée les plus explicites possibles pour que le lecteur puisse suivre les résultats présentés (*op. cit.* : 59). En effet, les processus de la catégorisation, du raisonnement et de l'explication se basent, aussi bien dans l'analyse quantitative que qualitative, sur les mêmes principes, même si les opérations dans l'analyse quantitative sont plus précises grâce à leur nature, c'est-à-dire grâce à un corpus bien délimité (*cf.* Mäkelä 1990 : 45).

maamut@utu.fi

BIBLIOGRAPHIE

ADAM J-M. (1990) : *Eléments de linguistique textuelle. Théorie et pratique de l'analyse textuelle* [deuxième éd.]. Liège, Mardaga.

ARNAUD, P. J. L. (1984) : « The lexical richness of L2 written productions and the validity of vocabulary tests », in T. CULHANE, C. KLEIN BRALEY & D. K. STEVENSON (eds.) : *Practice and problems in language testing, occasional papers 29*, Department of Language and Linguistics, Essex, University of Essex, 14-28.

ARNAUD, P. J. L. (1993) : « Estimations subjectives des fréquences des mots en L1 et en L2 », dans P. J. L. ARNAUD & P. THOIRON (dirs.) : *Aspects du vocabulaire*. Lyon, Presses universitaires de Lyon, 7-16.

BESSONNAT D. (1988) : « Le découpage en paragraphes et ses fonctions », *Pratiques 57*, 81-105.

CORDER S. P. (1980) : « Dialectes idiosyncrasiques et analyse d'erreurs », *Langages 57*, 17-28, [originellement publié dans IRAL, vol. IX/2., 1971].

COSSETTE A. (1994) : La richesse lexicale et sa mesure. *Travaux de linguistique quantitative 53*, Paris, Honoré Champion Éditeur.

DE PIETRO, J. F., MATTHEY, M. & PY, B. (1988) : « Acquisition et contrat didactique : séquences potentiellement acquisitionnelles dans la conversation exolingue », dans Actes du 3e Colloque Régional de Linguistique, Strasbourg.

Dictionnaire des fréquences (1971), Etudes statistiques sur le vocabulaire français, vocabulaire littéraire des XIX et XX siècles, quatre tomes, Nancy, C.N.R.S. – T.L.F.

DUGAST D. (1980) : *La statistique lexicale*. Genève, Editions Slatkine.

ELO A. (1993) : *Le français parlé par les étudiants finnophones et suédophones*. Annales Universitatis Turkuensis, série B, vol. 198, Turku, Université de Turku.

FAYOL M. (1989) : « Une approche psycholinguistique de la ponctuation : étude en production et en compréhension », *Langue française 81*, 21-39.

[13] Notre traduction.

FAERCH, C., HAASTRUP, K. & PHILLIPSON, R. (1984) : *Learner Language and Language Learning*. København, Gyldendalske Boghandel, Nordisk Forlag A.S.

Français fondamental (1er degré) (1972, première édition 1959). Paris, Publication de l'Institut National de Recherche et de Documentation Pédagogiques.

Français fondamental (2e degré), (s.d.), 2e édition, Paris, Publication de l'Institut Pédagogique National.

GAONAC'H D. (1984) : « La notion d'interlangue et la psychologie cognitive de langage », *in* B. PY (dir) *Acquisition d'une langue étrangère III*, Paris-Neuchâtel, Presses Universitaires de Vincennes et Centre de Linguistique Appliquée de Neuchâtel, 63-79.

GIACOBBE, J. (1992) : *Acquisition d'une langue étrangère : Cognition et interaction*. Paris, CNRS Editions.

GUSTAFSSON M. (1990) : « Lexical density as a style marker », in K. BATTARBEE & R. HILTUNEN (éds), *Alarums* & *Excursions :* Working Papers in English, University of Turku, Publications of the Department of English, 45-60.

LARSEN-FREEMAN, D. (1997) : « Chaos/complexity science and second language acquisition », *Applied Linguistics 18/2*, 141-165.

LINNARUD M. (1986) : *Lexis in Composition. A Performance Analysis of Swedish Learners' Written English*. Lund Studies in English 74, Malmö, GWK Gleerup, Liber Förlag.

LUNDQUIST L. (1980) : *La cohérence textuelle : syntaxe, sémantique, pragmatique*. København, Nyt Nordisk Forlag Arnold Busck.

LUNDQUIST L. (1983) : *L'analyse textuelle : méthode, exercices*. Paris, CEDIC.

MUTTA M. (1999) : *La compétence lexicale des étudiants finnophones en français. Etude sur la production écrite des apprenants*. Thèse de doctorat de 3ème cycle, Université de Turku, Département d'études françaises.

MUTTA, M. (à paraître) : « The impact of mixed context on lexical development – a case study of Finnish students learning French abroad », in A. HOUSEN & M. PIERRARD (éds) *Current Issues in Instructed Second Language Learning*. SoLA series, Brussels, Mouton de Gruyter.

MÄKELÄ K. (1990) : « Kvalitatiivisen analyysin arviointiperusteet », in K. MÄKELÄ (éd) *Kvalitatiivisen aineiston analyysi ja arviointi*. Helsinki, Gaudeamus, 42-61.

PÉRY-WOODLEY, M.-P. (1993) : *Les écrits dans l'apprentissage*. Paris, Hachette, coll. F.

PORQUIER, R. (1984) : « Communication exolingue et apprentissage des langues », dans B. PY (dir) : *Acquisition d'une langue étrangère III*, Paris-Neuchâtel, Presses Universitaires de Vincennes et Centre de Linguistique Appliquée de Neuchâtel, 17-47.

PORQUIER, R. (1986) : « Remarques sur les interlangues et leurs descriptions », *Etudes de Linguistique Appliquée 63*, 101-107.

PORQUIER, R. (1994) : « Communication exolingue et contextes d'appropriation : Le continuum acquisition /apprentissage », *Bulletin suisse de linguistique appliquée 59*, 159-169.

PRESTON D. R. & BAILEY R. (1996) : « Preface », in R. BAYLEY & D. R. PRESTON (éds.) *Second Language Acquisition and Linguistic Variation*, Amsterdam/Philadelphia, John Benjamins Publishing Company, xiii-xviii.

RIEGEL M., PELLAT J.-Ch. & RIOUL R. (1997) : *Grammaire méthodique du français*. 1ère éd. 1994, Paris, PUF.

SCHROEDER, M. (1995) : *Self-similarity : Chaos, fractals, power laws.* New York, WH Freeman.

SELINKER, L. (1975) : « Interlanguage », in J.-C. RICHARDS (ed) : *Error Analysis – Perspectives on Second Language Acquisition*, London, Longman [réimprimé à partir d'IRAL, vol. X/3 1972].

SUOMELA-SALMI E. (1997) : *Les syntagmes nominaux (SN) dans les discours économiques français : repères textuels*. Annales Universitatis Turkuensis, Série B, vol. 218, Turku, Université de Turku.

TARONE E. (1988) : *Variation in Interlanguage*. London, Edward Arnold.

VASSEUR, M.-Th. (1991) : « Solliciter n'est pas apprendre (initiative, sollicitation et acquisition d'une langue étrangère) », dans C. RUSSIER, H. STOFFEL & D. VÉRONIQUE (dirs.) *Interactions en langue étrangère*, Aix-en-Provence, Publications de l'Université de Provence, 49-59.

VOGEL, K. (1995) : *L'interlangue - la langue de l'apprenant*. Toulouse, Presses Universitaires du Mirail [Titre original (1990) *Lernersprache*, Tübingen, Gunter Narr Verlag].